PeMS Data Extraction Methodology and Execution Technical Memorandum

for the
Southern California Association of Governments

June 2006

Prepared by:

Urban Crossroads, Inc.
41 Corporate Park, Suite 300
Irvine, CA 92606

Table of Contents

Glossary of Terms

Imputation – The process of filling “holes” (voids) in the traffic detector data. Detectors can go bad from time to time and some detectors are bad all the time. This problem results in traffic data with a high number of holes. PeMS corrects these data holes using imputation algorithms.

G-factor – A combination of the average length of the vehicles in the traffic stream and the tuning of the loop detector itself. This value is used to estimate the average speed of traffic flow. Because the G-factor is a constant value and does not produce reliable speed estimates, PeMS computes this value on a regular basis.

PeMS – The Freeway Performance Measurement System (PeMS) is a real-time Archive Data Management System for transportation data. PeMS collects raw detector data in real-time, stores and processes this data, and reports this data through a number of web pages such that engineers can analyze the performance of freeway segments or of the overall freeway system.

Detector – A vehicle magnetic-sensor imbedded in the pavement and positioned within a travel lane such that vehicles traveling over the sensor are “detected” and reported to a roadside station.

Single Loop Detectors - A majority of the loop detectors in the State of California are single loop detectors. This means that there is only a single detector station in a lane. Speed cannot be measured directly from single loop detector and has to be estimated using the G-factor and the loop data values.

Double Loop Detectors - Double loop detectors are installed at specified spacing and thus have the capability of measuring vehicle speed. PeMS has found that the speed calculation algorithms based solely on the flow and occupancy generated by PeMS are better than the speed measurements from the double loops so the double loop speed measures are not used in the PeMS.

Flow – The number of vehicles that crossed the detector during a given time period

Speed – The PeMS reported average speed of all vehicles passing a given loop or series of loops.

Occupancy - The percentage of time that the detector was occupied (in the “on” state) during a given time period.

Introduction

As an element of its regional freeway traffic model improvement and validation process, the Southern California Association of Governments (SCAG) is expanding the number of locations where model results are compared with reliable observed hourly traffic counts. To fulfill this objective, hourly freeway traffic volumes and speeds have been extracted from the California Department of Transportation's (Caltrans) Freeway Performance Measurement System (PeMS). This technical memorandum discusses the methodologies and tools used to extract and summarize the hourly traffic data obtained from PeMS.

PeMS Overview

The Freeway Performance Measurement System (PeMS) is a project to investigate various performance measures on the freeway system. It is a joint effort by Caltrans, the University of California, Berkeley, and Path, the Partnership for Advanced Technology on highways.

The PeMS 6.2 is a software/database tool designed at UC Berkeley to host, process, retrieve and analyze road traffic conditions. The PeMS database logs data from California freeway traffic detectors, as well as incident-related data from the California Highway Patrol (CHP) and associated weather data. PeMS features a web-based Graphical User Interface (GUI) that provides the ability to easily extract various representations of the data.

Data processed in PeMS is handled by special algorithms that run in the background and populate the processed tables. Detection and imputation of data errors are crucial first steps in the data flow. Bad and missing samples present problems for any algorithm that use the data for analysis. PeMS expends considerable effort detecting bad data from the measurements and bad detectors themselves.

Currently PeMS collects, process, stores and makes available only data from the six Caltrans Districts (D3, D4, D7, D8, D11, D12). These key districts include the six (6) major urban areas in California.

Data Extraction Methodology

Many of the existing methodologies developed by the PeMS developers, such as the calculation of speed, have been employed in the data extraction process. The methodologies leveraged include:

1. The aggregation of the 30-second raw data to 5-minute then hourly data samples.
2. The imputation of data.
3. The calculation of speeds.
4. The estimation of truck volumes.

The following sections discuss the PeMS methodologies. For more detailed information on PeMS, please refer to the numerous technical white papers published since the origination of PeMS and available through PeMS 6.3 and the University of Berkeley.

Collection and Aggregation

The primary function of PeMS is to process traffic detector data. PeMS receives measurements (flow and occupancy) from approximately 28,000 individual lane detectors and reports these measurements every 30 seconds. PeMS data is then reduced to a set of numbers that represent the performance of the freeway system in the entire state for a given year. The data processing steps for the freeway data collected include:

1. PeMS receives the raw, 30-second data from each of the reporting Caltrans districts. As not all data is received for the individual districts in the same data format, the data is then transformed into the PeMS schema and archived in the PeMS database.
2. The 30-second data is aggregated to 5-minute samples per lane. This data is referred to as the 5-minute data with holes.
3. The 5-minute data with holes is then imputed. The resultant new set of 5-minute data without holes is uploaded into the database.
4. The 5-minute aggregate data set is then consolidated to hourly aggregate data samples.

The G-factors are computed for each detector in the system. The G-factors for an area are stored in the database and are used to compute speed in real-time.

Imputation parameters are calculated offline from the 5-minute data with holes. The imputation parameters, like the G-factors, are stored in the database and aren't computed in real-time.

Each detector has a diagnostics performed on it every day. PeMS makes a determination as to whether a detector is good or bad. This determination is stored in the PeMS database for ongoing evaluation of individual detectors. Values are imputed for each of the bad detectors using the imputation parameters stored in the database.

Loop Detector Health

PeMS devotes a great deal of effort in identifying bad detectors so that data from these detectors can be omitted from the data samples.

PeMS classifies two major types of detector contamination: 1) detectors with missing data samples and 2) detectors with erroneous data samples. Each 24-hour period, a loop should report 24x60x2 = 2880 samples. In October 31, 2003, the sample-reporting rate in different districts varied between 50 and 90 percent. PeMS calculates the reporting rate by counting the number of samples it receives.

A reported sample could still be erroneous because of contaminated or otherwise "malfunctioning" detectors. To determine whether a loop detector is malfunctioning,

PeMS conducts an elaborate statistical test based on a time series of samples from the same loop detector.

There are six (6) tests that are performed on the data at the end of each day to help determine if a detector is bad. Once a detector is determined as bad, data for that detector is imputed based on a set of complex algorithms outlined below.

Imputation Techniques

Imputation is the process of filling in for holes in the detector data. A hole can come about for two reasons: 1) data for that detector was not received for a particular time period, or 2) data was received, but it is believe that the data is bad as indicated by the diagnostic routines. In either case, a series of imputation methods are applied in an attempt to fill in the missing data.

Different imputation routines require different amounts of data. The process that is run depends on the amount of data for each bad detector. The imputation routines used in PeMS are applied in the order below.

- ***Linear Regression from Neighbors based on Local Coefficients***
  This algorithm imputes the missing values at a detector based on the surrounding neighbors. A neighbor is a loop in other freeway lanes at the same location as well as all of the loops at the locations immediately upstream and downstream of the current loop.

- ***Linear Regression from Neighbors based on Global Coefficients***
  This algorithm is similar to the one above but in situations where loops never report data it is not possible to compute local regression coefficients. Therefore global regression coefficients that represent the general relationship seen throughout the district are used.

- ***Temporal Medians***
  If adjacent detectors do not report data, linear regression does not work. In this case if you have a hole at 3pm on Wednesday at a particular loop then you can look back over the past N Wednesdays at 3pm for this loop and take the median of those measurements.

- ***Cluster Medians***
  This is the last method of imputation. A cluster is a group of detector stations that have the same macroscopic flow and occupancy patterns over the course of a typical week.

  A station is assigned to a cluster and any holes are assigned to the cluster ID. If all of the other imputation methods fail then the quantity profiles for the cluster are pulled out of the database and the correct value for this day of the week and time of day is assigned.

Speeds and G-factors

PeMS has the ability to compute speed for sensors that don't report speed, such as single loop detectors. For sensors reporting speed, like double loop detectors, speed measurements can be used.

Traditionally speed is calculated using the G-factor. The G-factor is a value that represents the effective length of the vehicle and the tuning of the loop detector. When calculating speed, PeMS does not use a constant value[1] for the G-factor. Instead, PeMS estimates a G-factor for each loop for every 5 minutes over an average week to provide accurate G-factors to estimate speeds.

The speed algorithm used in PeMS is adapted from "Statistical Methods for Estimating Speed Using Single Loop Detectors" by van Zwet, Chen, Jia and Kwon. The algorithm has been tested and validated against data collected from double loop detectors and floating cars.

The steps for estimating speed are as follows:

1. It is assumed that speed on the freeway at free-flow conditions is known and constant[2]. Starting backwards, the G-factor is computed for a number of points during a number of days.
2. An adaptive regression method is used to obtain a G-factor for each loop in the system over a typical week.
3. The G-factor is used along with the loop data to estimate speed.

Estimating Truck Volumes

PeMS has the ability to estimate truck volumes where detectors do not report truck flow. The algorithm attempts to break down the total flow into passenger cars and large trucks. The following assumptions are applied in the algorithm:

1. There is no truck flow in Lane 1 (the inside or median lane).
2. For freeways that are multi-lane, the vehicle speeds over different lanes are synchronized.
3. The traffic volume consists mostly of short passenger cars and long trucks.
4. The average length of passenger cars is 16 feet and the average length of trucks is 60 feet.

[1] Typically, a constant value for the g-factor is used. Because g-factor varies by lane, time-of-day, as well as the loop sensitivity, inaccurate speeds will be calculated if a constant is used.

[2] The constant values have been verified using double loops in the Bay area to measure actual free flow speeds on different types of freeways and in different lanes. The constant values can be localized (i.e., fine-tuned) to give a greater estimate in other geographical areas.

The algorithm estimates the proportion of trucks in each lane, starting with the first lane, Lane 1, and working to the outer lanes. PeMS stores the proportion of trucks in each lane. The data is usually plotted as an aggregate across all lanes because individual per-lane estimates tend to be noisy.

It is important to note that the truck volume estimates from the PeMS algorithm are not measured but estimated from 5-minute aggregated volume and occupancy data. It is not recommended to look at the truck volume estimate from a single location over a single day. A better way to use the PeMS truck volume estimates is to look at the outputs over many locations, preferably along the corridor in question over multiple days (preferably over 5 weekdays).

Tool Execution

Numerous tasks were performed to extract and summarize the hourly flow and speed data from PeMS. The following sections describe the extraction and summarization processes as well as tools developed and utilized to fulfill the requirements of the project.

Extraction

Using a PeMS database account, area cities and counties, detector characteristics, hourly volume and speed data were extracted from the database using the following SQL queries. These queries were executed on the Oracle database server and exported to ASCII format on the server. Once the extraction process was complete, the ASCII files were downloaded to a local workstation for additional manipulation (i.e., summarization).

Query 1) getCC.sql - Acquires a list of the cities and counties stored in the PeMS database. As only coded domain values are used for the city and county fields in the station configuration table in the PeMS database, the city and county identifiers and names were needed to replace the cryptic values in the query results (Query 2 and Query 3).

```
SET TERMOUT OFF ;
SET HEADING OFF ;
SET LINESIZE 400 ;
SET PAGESIZE 0 ;
SET ECHO OFF ;

SPOOL Cities ;
SELECT * FROM pems6.cities ;
SPOOL OFF ;

SPOOL counties ;
SELECT * FROM pems6.counties ;
SPOOL OFF ;

EXIT ;
```

Query 2) getVdsMl.sql - Acquires station configuration characteristics for all mainline VDS located in Caltrans Districts 7, 8 and 12.

```
SET TERMOUT OFF ;
SET HEADING OFF ;
SET LINESIZE 200 ;
SET PAGESIZE 0 ;
SET ECHO OFF ;
SPOOL ScagVdsHv ;

SELECT FREEWAY_ID, FREEWAY_DIR, NAME, COUNTY_ID, CITY_ID,
      STATION_ID, STATION_TYPE, PHYSICAL_LANES,
      STATE_POSTMILE, ABS_POSTMILE, LATITUDE, LONGITUDE, LENGTH
  FROM pems6.station_config_current_vw
    WHERE STATION_TYPE = 'ML'
    AND DISTRICT_ID IN (7,8,12)
  ORDER BY FREEWAY_ID, FREEWAY_DIR, ABS_POSTMILE ;
SPOOL OFF ;

EXIT :
```

Query 3) getVdsHv.sql – Acquires station configuration characteristics for all HOV VDS located in Caltrans Districts 7, 8 and 12.

```
SET TERMOUT OFF ;
SET HEADING OFF ;
SET LINESIZE 200 ;
SET PAGESIZE 0 ;
SET ECHO OFF ;
SPOOL ScagVdsHv ;

SELECT FREEWAY_ID, FREEWAY_DIR, NAME, COUNTY_ID, CITY_ID,
      STATION_ID, STATION_TYPE, PHYSICAL_LANES,
      STATE_POSTMILE, ABS_POSTMILE, LATITUDE, LONGITUDE, LENGTH
  FROM pems6.station_config_current_vw
    WHERE STATION_TYPE = 'HV'
    AND DISTRICT_ID IN (7,8,12)
  ORDER BY FREEWAY_ID, FREEWAY_DIR, ABS_POSTMILE ;
SPOOL OFF ;

EXIT ;
```

Query 4) getHr1.sql - Acquires traffic data, such as flow, occupancy and speed, from the hourly summary table in the PeMS database. Data was extracted for a time period starting on January 1, 2003 through December 31, 2003. A batch file (GetHRVol1.bat) executes Query 4 for each mainline and HOV VDS obtained from the execution of Query 2 and Query 3. The '&1' variable is used as the placeholder for each VDS as the batch file looped through the VDS results list.

```
SET TERMOUT OFF ;
SET HEADING OFF ;
SET LINESIZE 240 ;
SET PAGESIZE 0 ;
SET ECHO OFF ;
SPOOL &1
SELECT TO_CHAR(time_id, 'mm-dd-yyyy hh24:mi') Time, STATION_ID,
      TOT_FLOW, TOT_TRUCK_FLOW, AVG_OCC, AVG_SPEED,
      TOT_SAMPLES, LANE_PTS, LANE_PTS_OBS,
      LANE_PTS_METHOD_1, LANE_PTS_METHOD_2,
      LANE_PTS_METHOD_3,LANE_PTS_METHOD_4
  FROM pems6.station_hour_summary
    WHERE time_id BETWEEN TO_DATE('01/01/2003 00:00:00',
                          'mm/dd/yyyy hh24:mi:ss')
                AND TO_DATE('12/31/2003 23:59:59',
                          'mm/dd/yyyy hh24:mi:ss')
                AND STATION_ID = &1;
SPOOL OFF ;

EXIT ;
```

Summarization

The Mathworks' MATLAB analysis software (version 2006a) was used to summarize the data extracted from the PeMS database. MATLAB is a high-level technical computing language and interactive environment for algorithm development, data visualization, data analysis, and numeric computation.

Three (3) programs were developed in MATLAB to summarize the extracted PeMS data. These programs are: 1) CallLoadVolumesVo1.m, 2) LoadHrVols02.m and 3) Vols2Excel.m. Listed below is a description of each. The MATLAB files used in the PeMS Data Extraction Project have been included as deliverables of the contract. SCAG may choose to license the MATLAB software to re-summarize the traffic data extracted from the PeMS database.

Program 1) CallLoadVolumesV01.m – Retrieves the information stored in ASCII text files. This program calls LoadHRVols02.m.

Program 2) LoadHrVols02.m - Summarizes and creates matrices of the hourly volume and speed data extracted from the PeMS database. The data is summarized by:

- Total Flow Weekdays
- Total Truck Flow Weekdays
- Total Speed Weekdays
- Total Flow Saturdays
- Total Truck Flow Saturdays
- Total Speed Saturdays
- Total Flow Sundays
- Total Truck Flow Sundays
- Total Speed Sundays

Program 3) Vols2Excel.m - Further summarizes data in the MatLab matrices and exports the data results to Microsoft Excel. The following information is provided in the Excel spreadsheets:

- **VDS-Summary** - Detector information such as facility, dir, VDS, location, longitude., latitude., ADT (week, Sat, Sun), L-Factors, Flows by 4 time periods, Flows by AM and PM peak hour, average speeds by four time periods, average speeds by AM and PM peak hour.
- **MeanFlowWk** - Weekday flows by hour of day (0-1, 1-2, ... 23-24) for each detector.
- **MeanTrkFlowsWk** - Weekday truck flows by hour.
- **MeanFlowsWk**nSpeedsWk - Weekday average speeds by hour
- **MeanFlowsSa** - Saturday flows by hour.
- **MeanTrkFlowsSa** - Saturday truck flows by hour.
- **MeanSpeedsSa** - Saturday average speeds by hour.
- **MeanFlowsSu** - Sunday flows by hour.
- **MeanTrkFlowsSu** - Sunday truck flows by hour.
- **MeanSpeedsSu** - Sunday average speeds by hour. FlowCountSa - Number of days, data collected on Saturdays. Maximum days = 52.
- **FlowCountWk** - Number of days, data collected on weekdays. Maximum days = 260.
- **FlowCountSa** - Number of days, data collected on Saturdays. Maximum days = 52.
- **FlowCountSu -** Number of days, data collected on Sundays. Maximum days = 52.

A variable (propObsT) has been set in this program. This variable represents the proportion of points that must be observed (i.e., real vs. imputed data) to be determined as "good" data. The observed rate is calculated in PeMS when the raw 30-second data is aggregated to 5-minute samples. Currently the propObsT variable is set to 50 percent. A 50 percent observed rate means that at a minimum 6 of 12 points must be observed within an hour time period to be determined as "good" data. With the purchase of the MATLAB software, SCAG may choose to update the propObsT variable and then rerun the data summaries based on the new interpretation rate of "good" data.

Validation

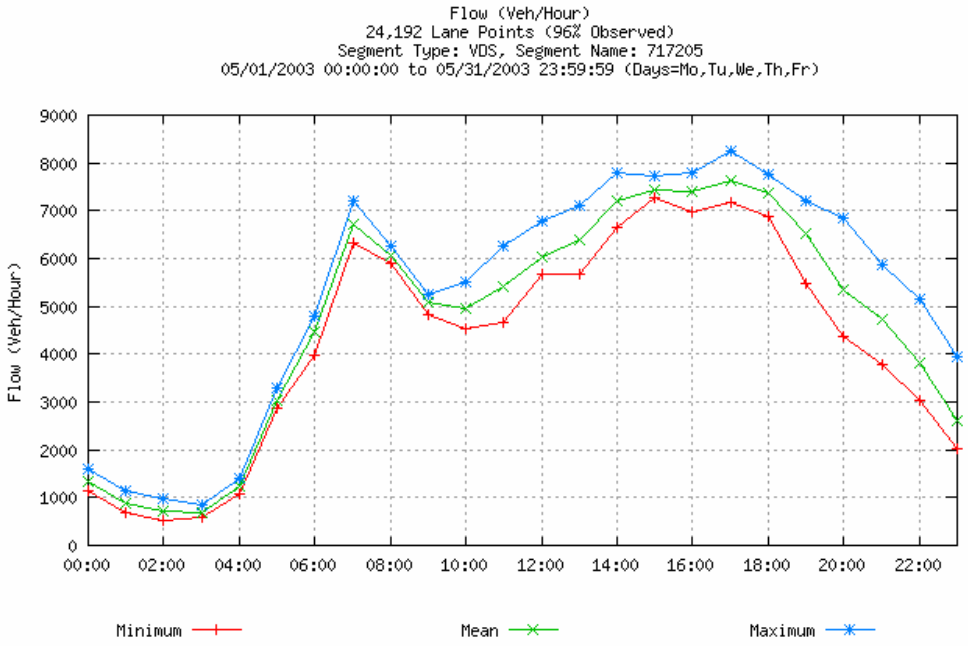
Ten (10) mainline VDS were identified for validating the data extraction tool. The VDS selected are matched locations with high observed rates. The charts presented below were generated in MATLAB using data for 251 non-holiday weekdays.

Overall, the profiles of the data extracted from the PeMS database compared to the data queried on the PeMS website are identical. Small discrepancies between the two are caused by the time limitation in the PeMS website application (i.e., the maximum period of time that can be queried is one (1) month).

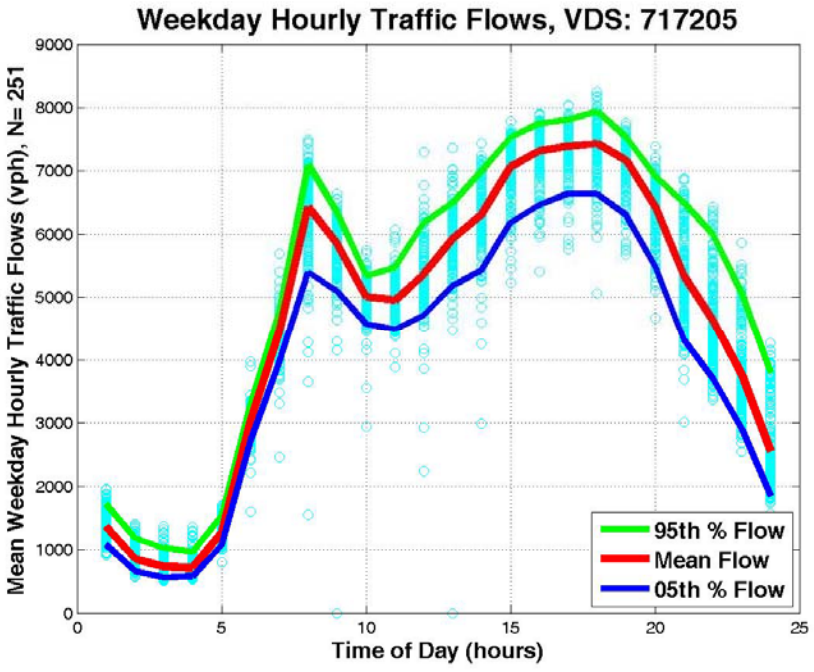
Sample 1 (Flow): Mainline #717205

Caltrans District:	7	CA Post Mile:	37.59
County:	Los Angeles	Absolute Post Mile:	36.090
City:	West Covina	Length:	0.545
Freeway:	I-10 E	Latitude:	34.0719
Name:	Citrus 2	Longitude:	-117.8884

PeMS Website

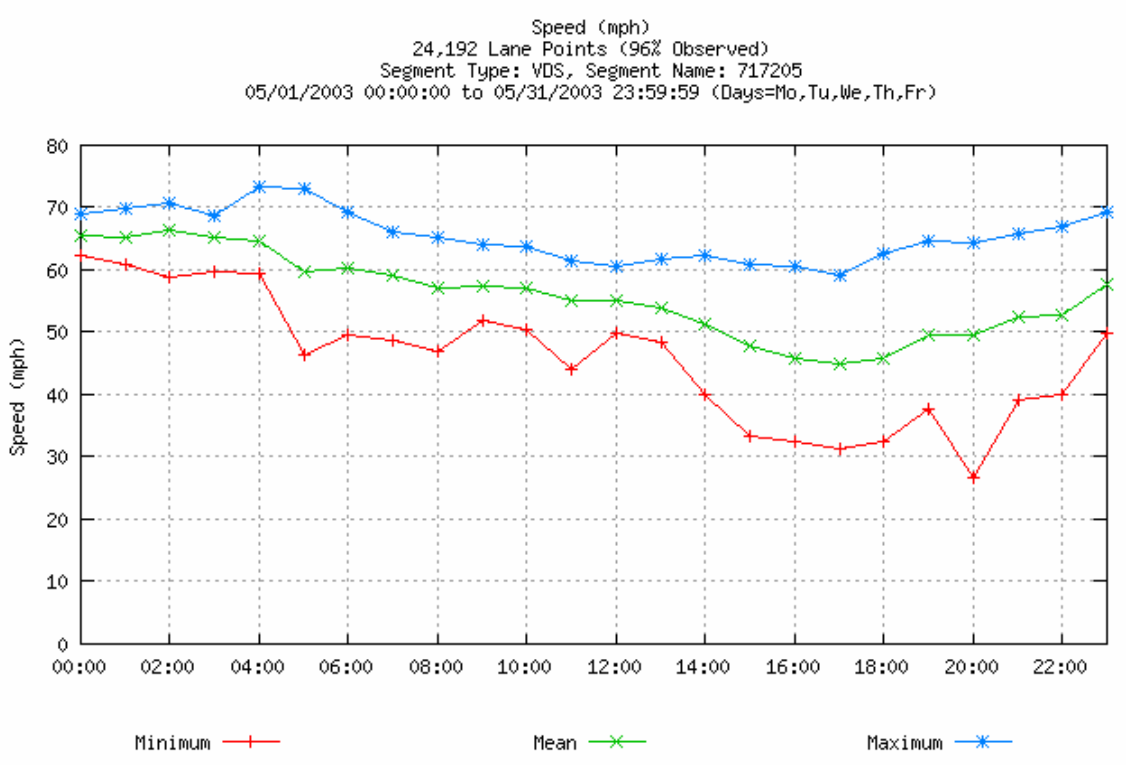


Extracted Dataset (MATLAB)

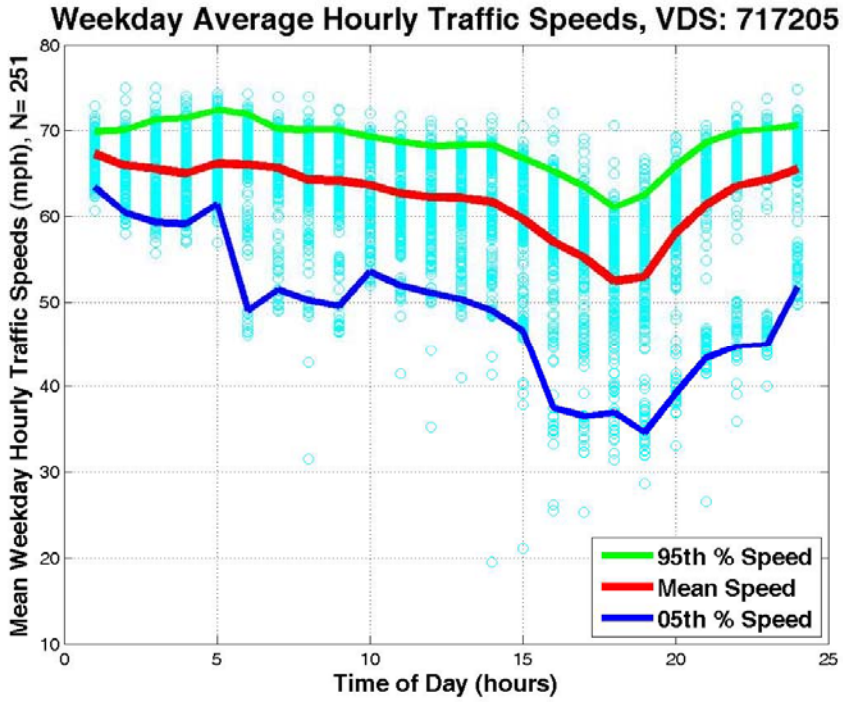


Sample 1 (Speed): Mainline #717205

PeMS Website



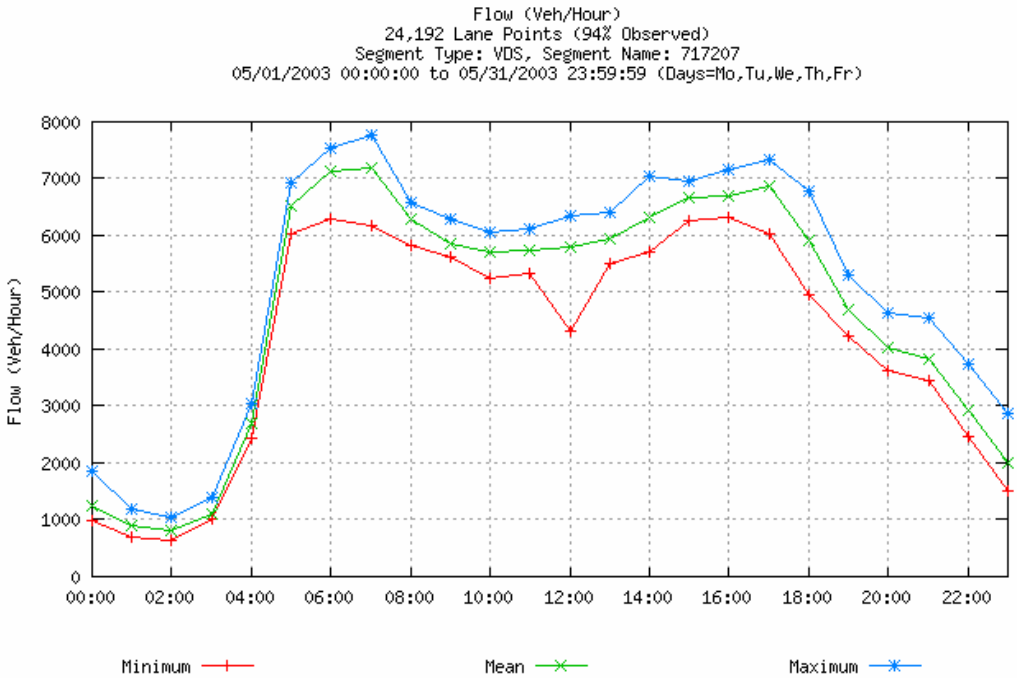
Extracted Dataset (MATLAB)



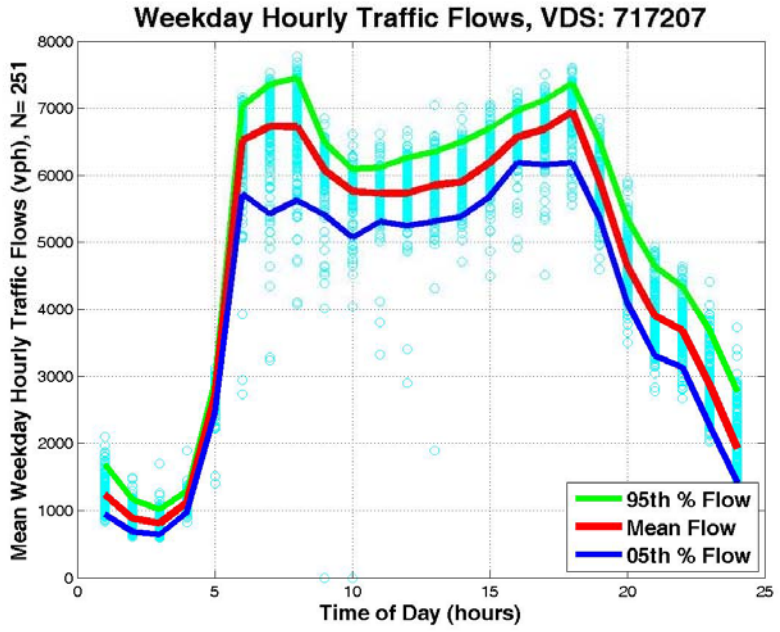
Sample 2 (Flow): Mainline #717207

Caltrans District:	7	CA Post Mile:	37.52
County:	Los Angeles	Absolute Post Mile:	36.02
City:	West Covina	Length:	0.245
Freeway:	I-10 W	Latitude:	34.0719
Name:	Citrus 2	Longitude:	-117.8896

PeMS Website

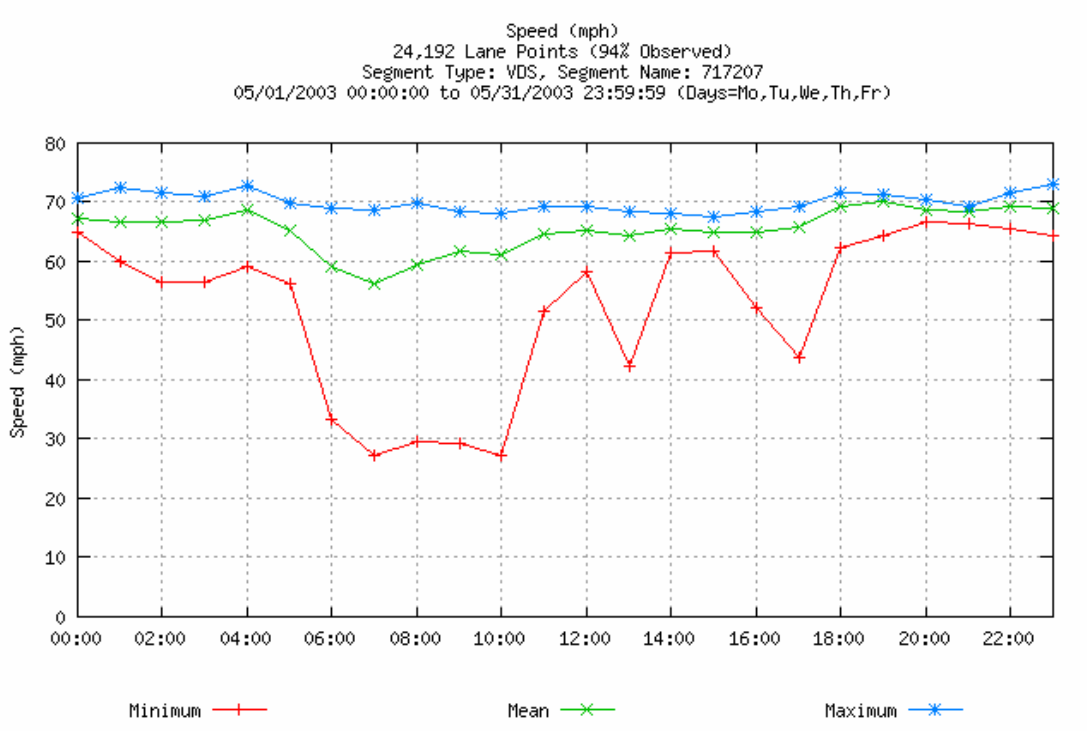


Extracted Dataset (MATLAB)

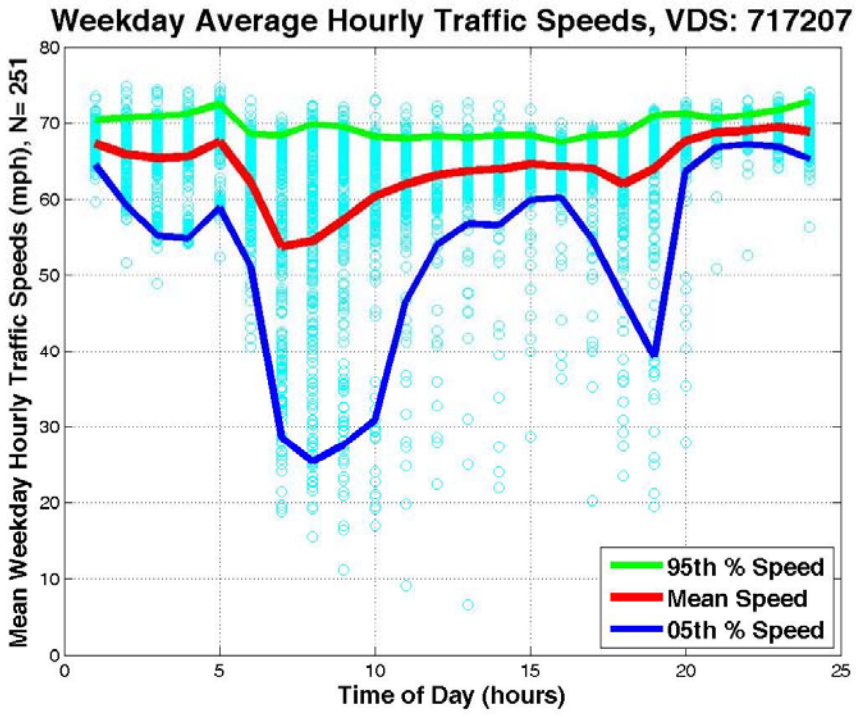


Sample 2 (Speed): Mainline #717207

PeMS Website



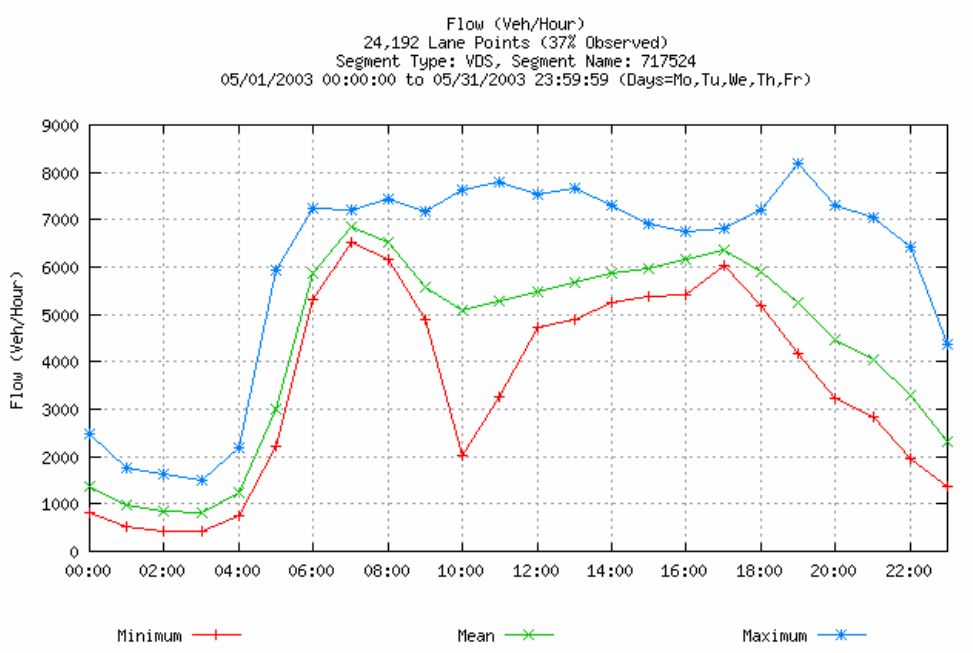
Extracted Dataset (MATLAB)



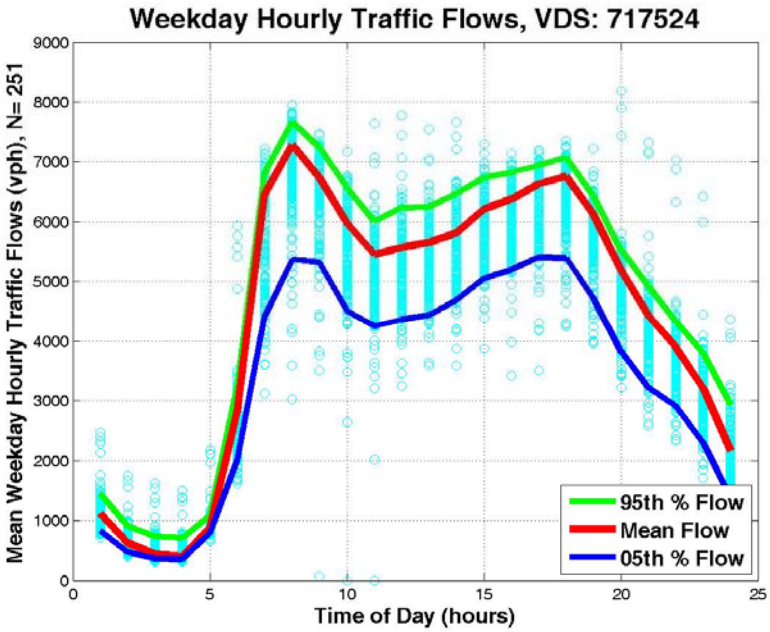
Sample 3 (Flow): Mainline #717524

Caltrans District:	7	CA Post Mile:	26.86
County:	Los Angeles	Absolute Post Mile:	28.219
City:	Los Angeles	Length:	1.030
Freeway:	101-N	Latitude:	34.1623
Name:	Woodlake	Longitude:	-118.6300

PeMS Website

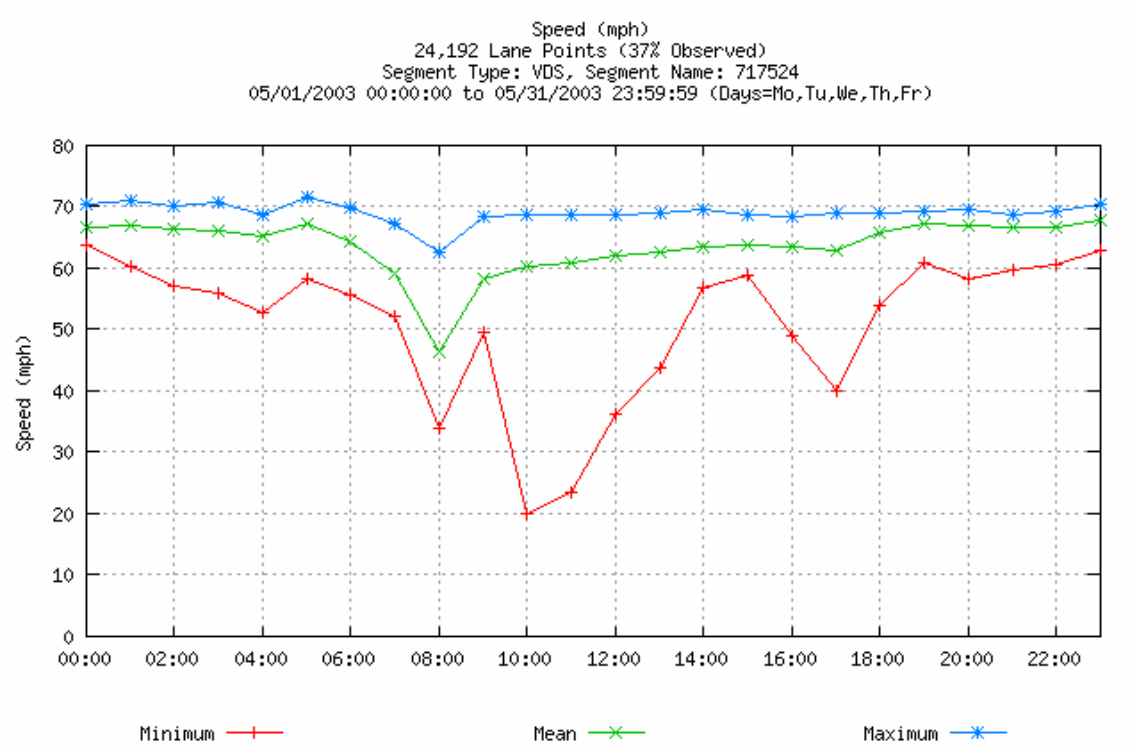


Extracted Dataset (MATLAB)

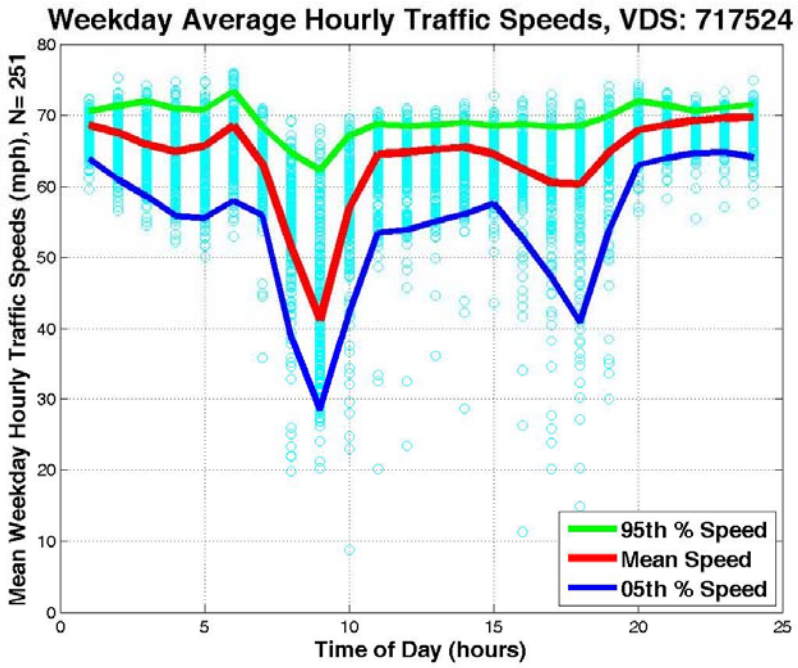


Sample 3 (Speed): Mainline #717524

PeMS Website



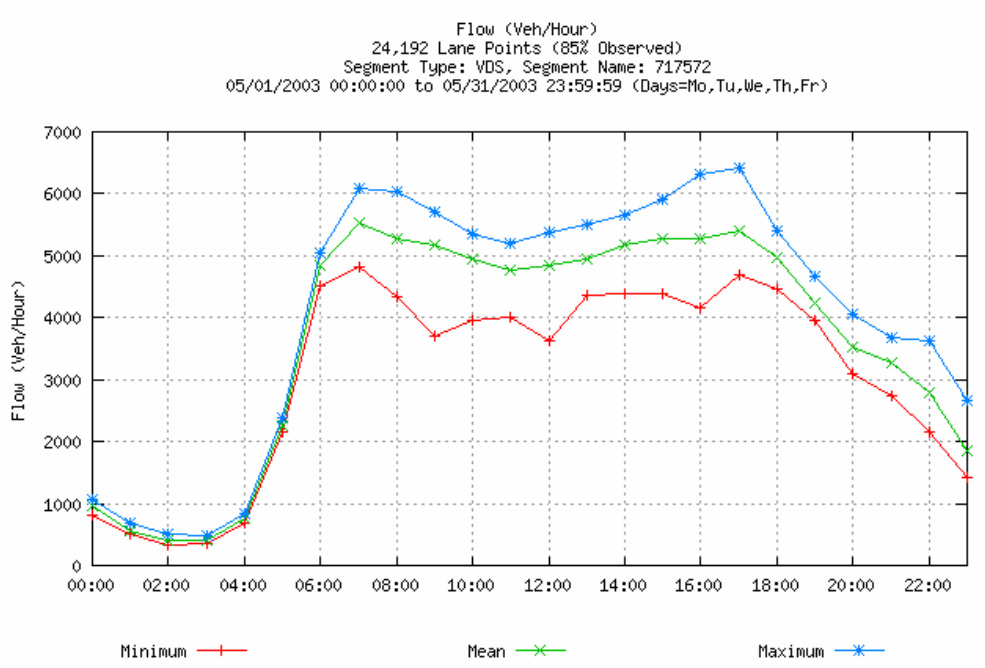
Extracted Dataset (MATLAB)



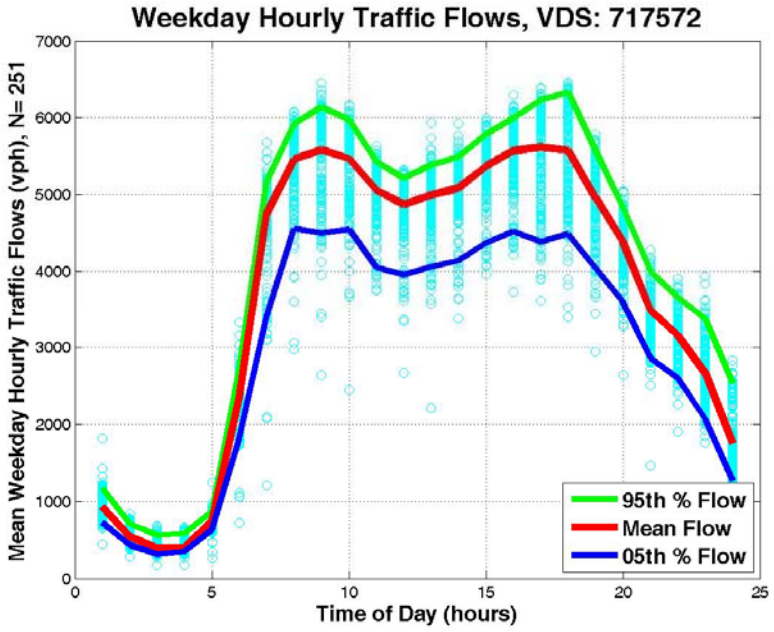
Sample 4 (Flow): Mainline #717572

Caltrans District:	7	CA Post Mile:	2.86
County:	Los Angeles	Absolute Post Mile:	2.860
City:	Burbank	Length:	0.915
Freeway:	134-W	Latitude:	34.1535
Name:	Buena Vista	Longitude:	-118.3270

PeMS Website

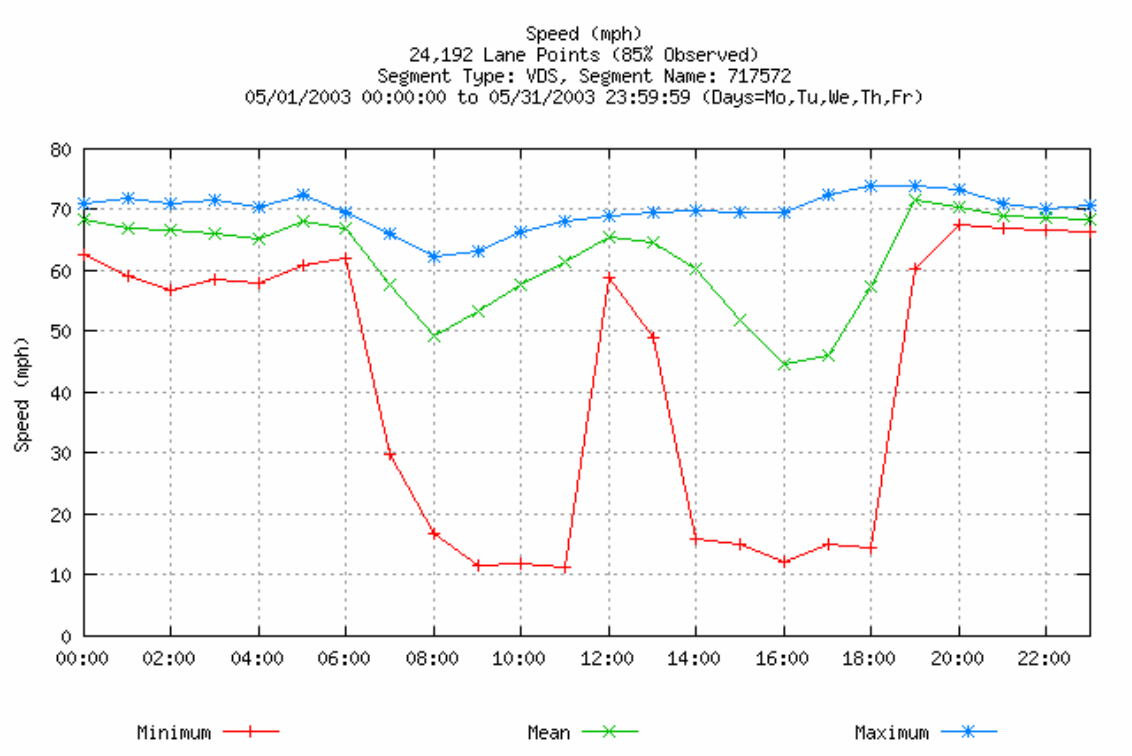


Extracted Dataset (MATLAB)

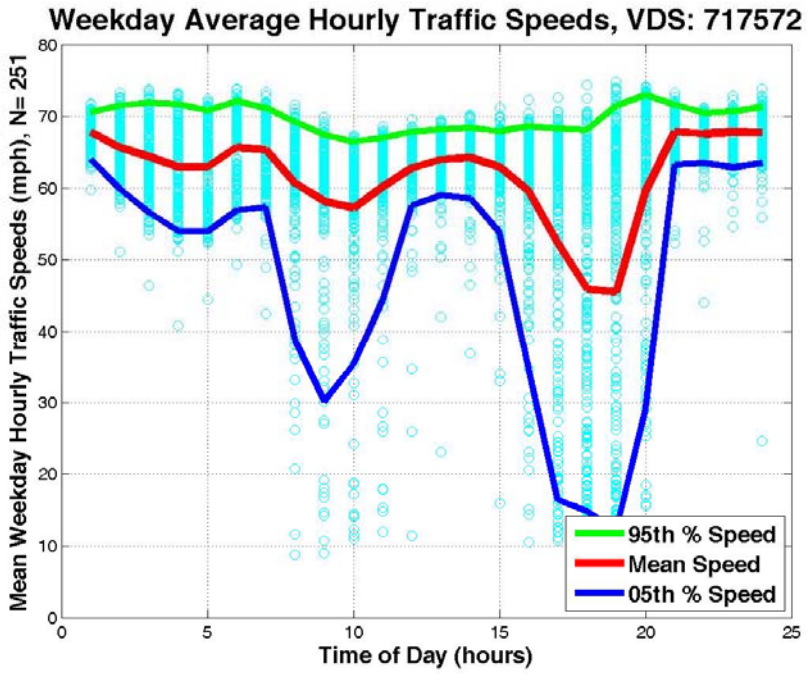


Sample 4 (Speed): Mainline #717572

PeMS Website



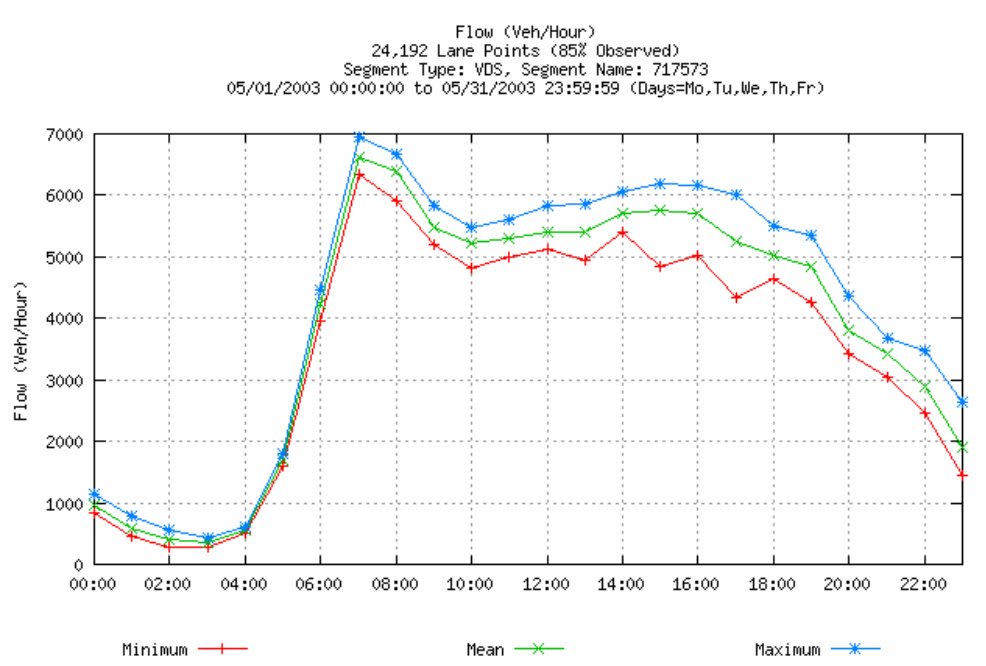
Extracted Dataset (MATLAB)



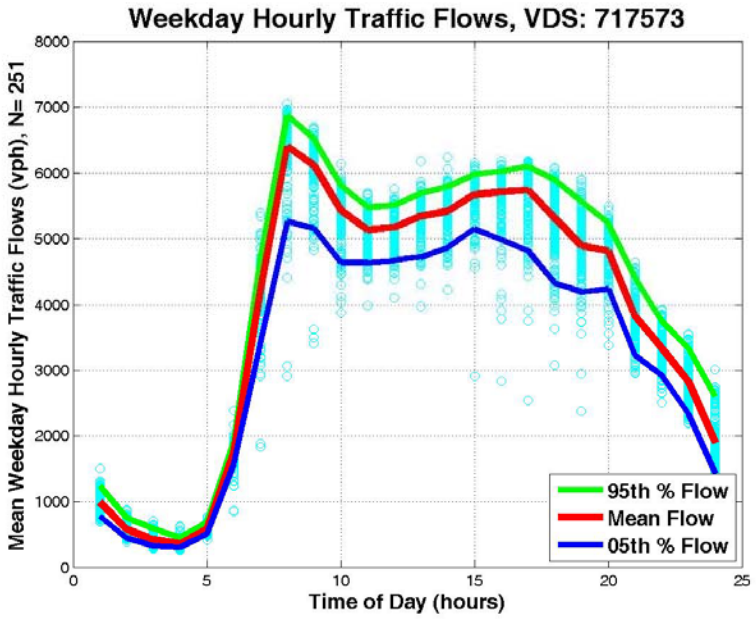
Sample 5 (Flow): Mainline #717573

Caltrans District:	7	CA Post Mile:	3.04
County:	Los Angeles	Absolute Post Mile:	3.040
City:	Los Angeles	Length:	0.885
Freeway:	134-E	Latitude:	34.1542
Name:	Buena Vista	Longitude:	-118.3235

PeMS Website

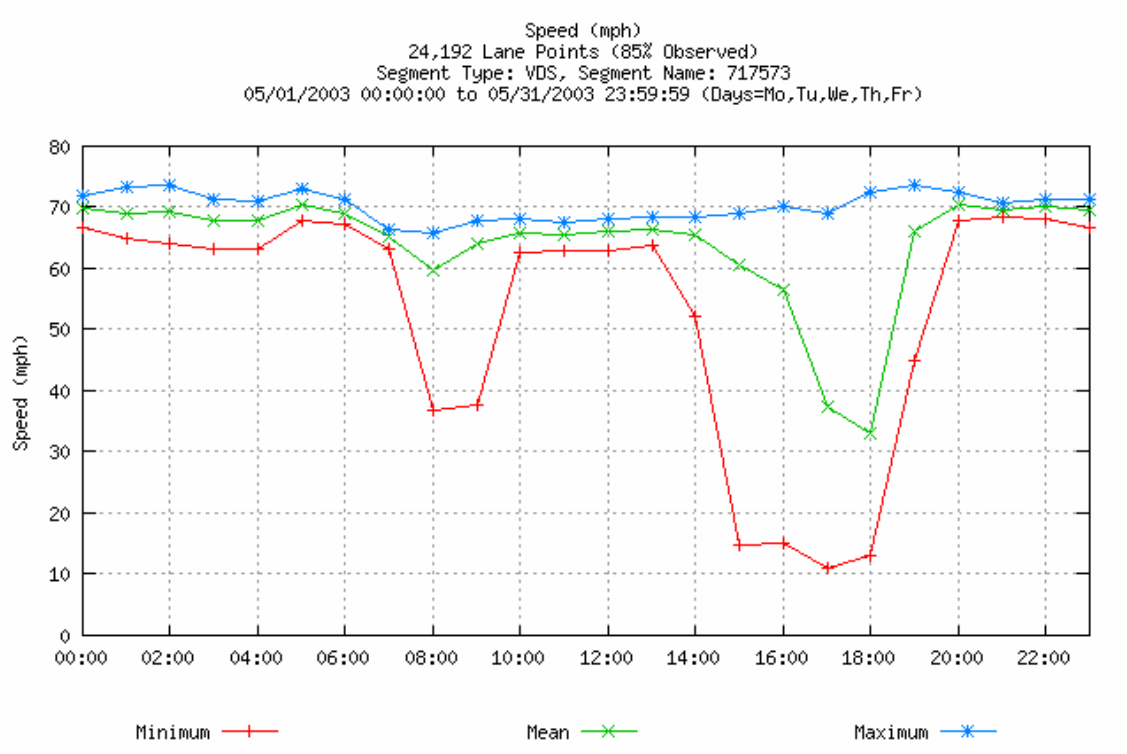


Extracted Dataset (MATLAB)

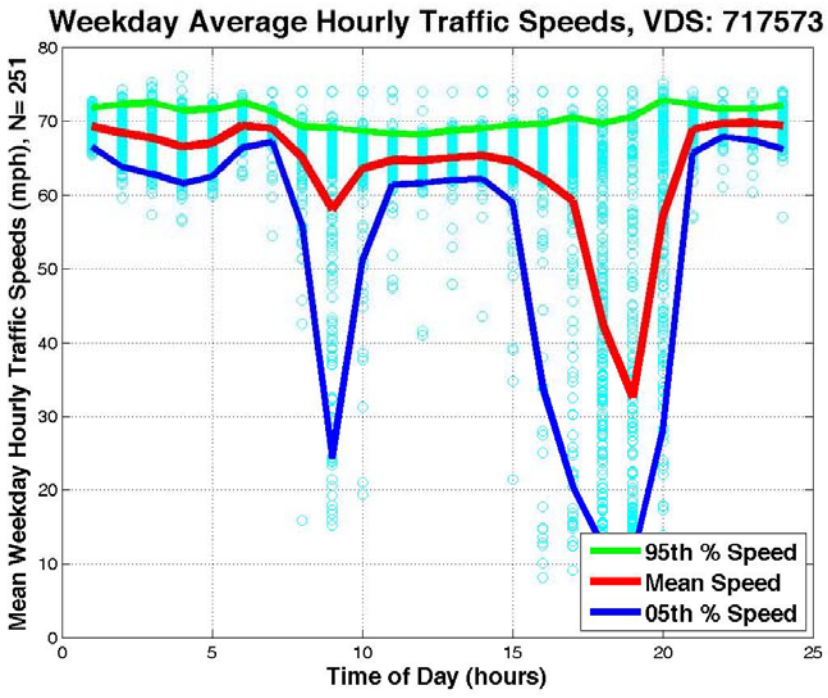


Sample 5 (Speed): Mainline #717573

PeMS Website



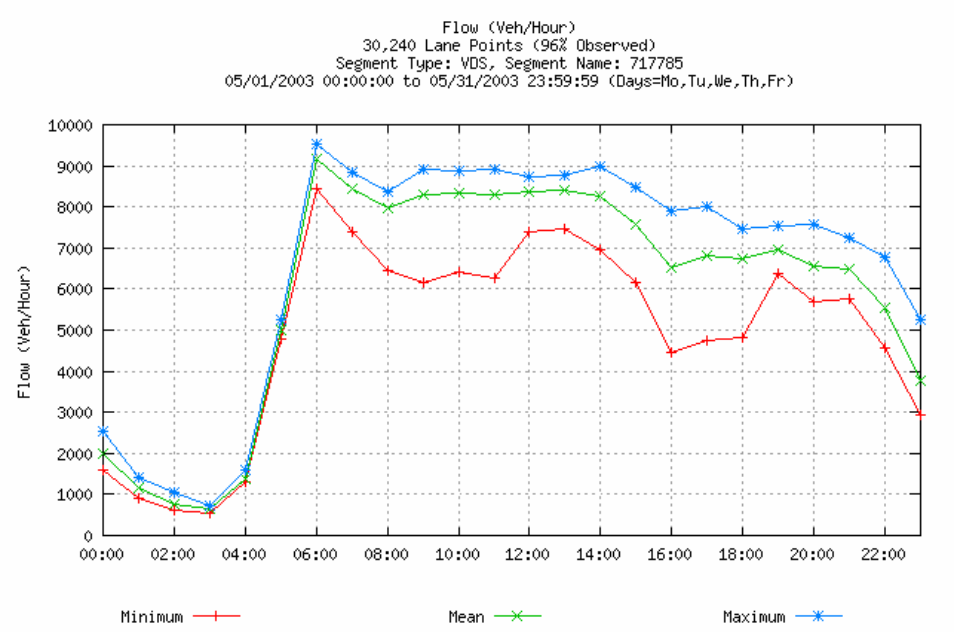
Extracted Dataset (MATLAB)



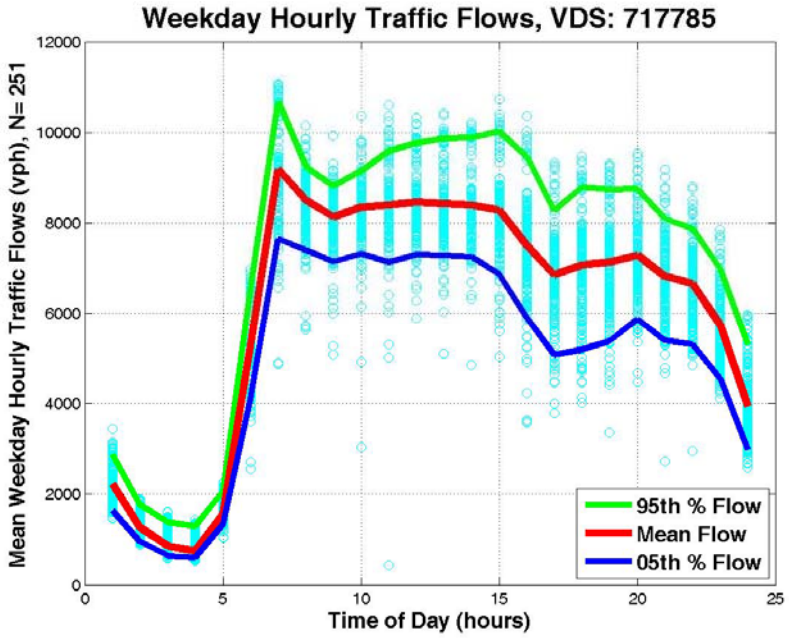
Sample 6 (Flow): Mainline #717785

Caltrans District:	7	CA Post Mile:	26.9
County:	Los Angeles	Absolute Post Mile:	50.672
City:	Los Angeles	Length:	.600
Freeway:	405-N	Latitude:	33.9993
Name:	Braddock	Longitude:	-118.4085

PeMS Website

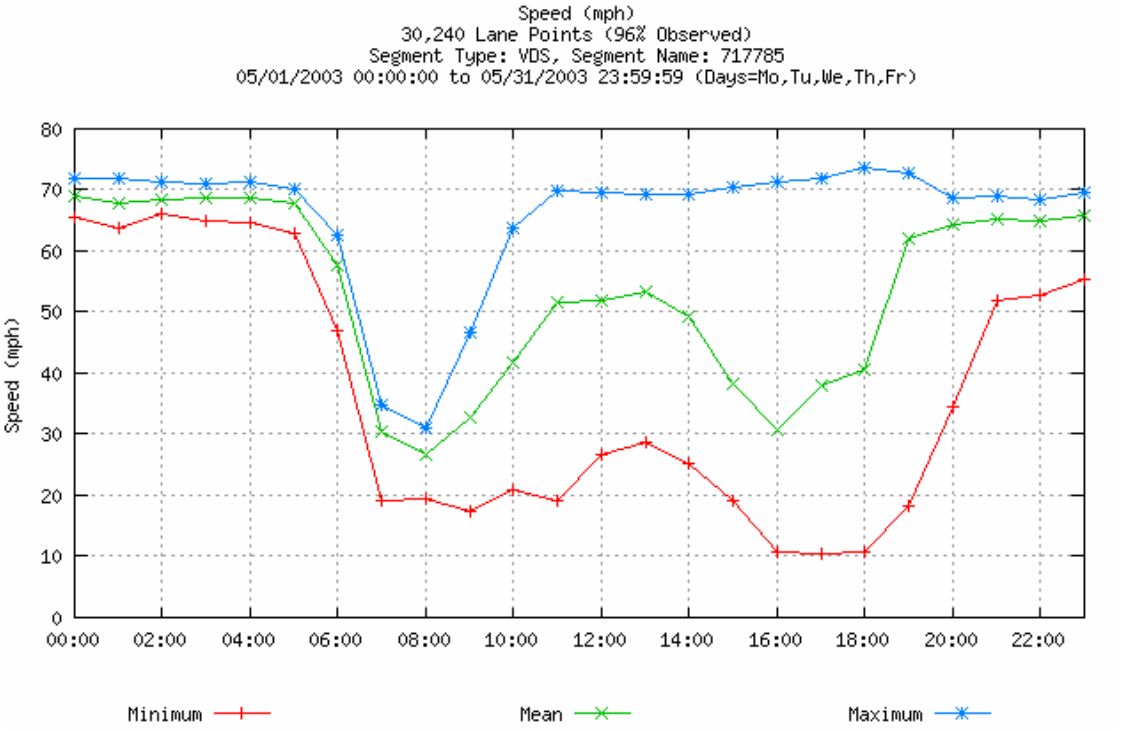


Extracted Dataset (MATLAB)

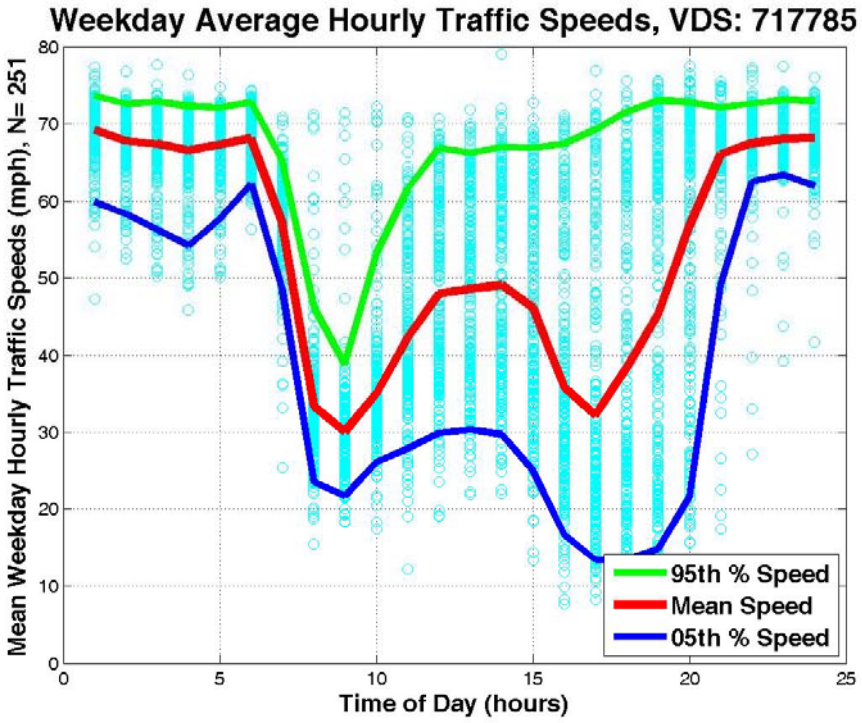


Sample 6 (Speed): Mainline #717785

PeMS Website



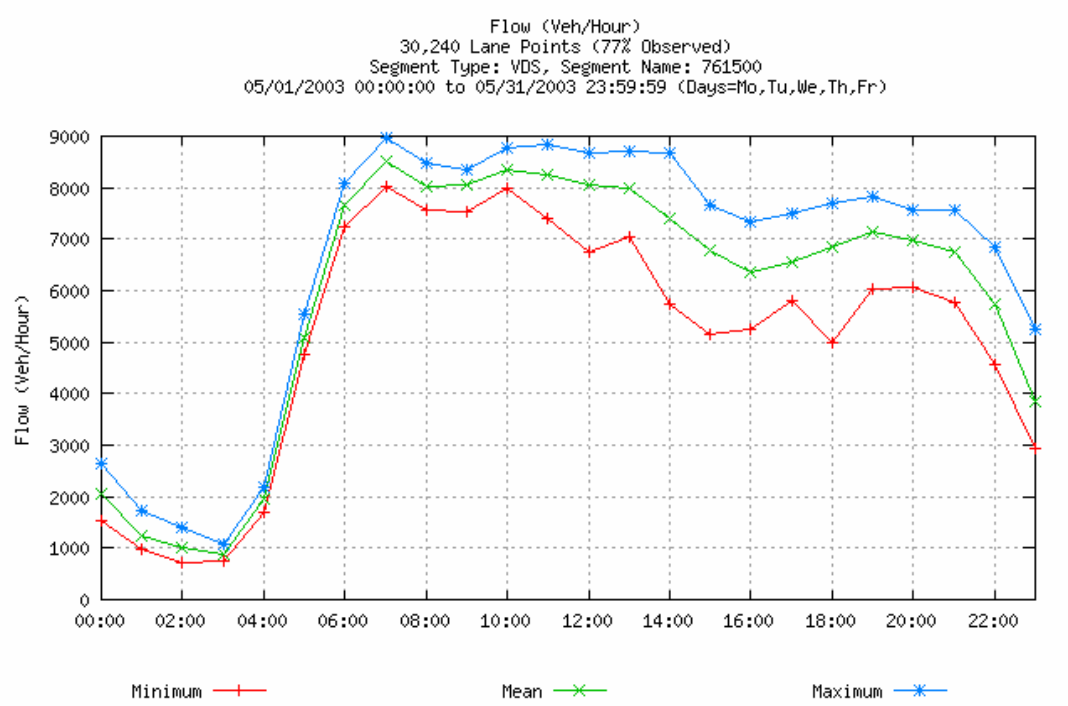
Extracted Dataset (MATLAB)



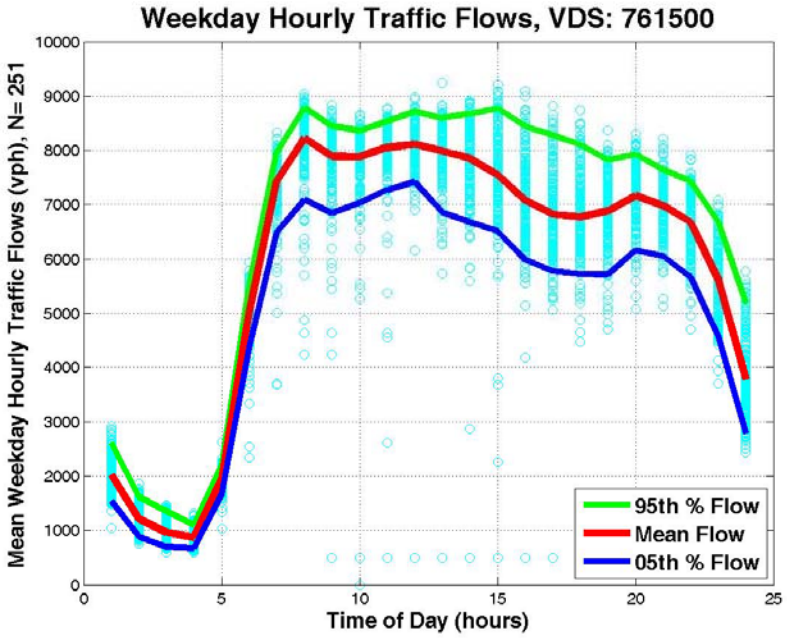
Sample 7 (Flow): Mainline #761500

Caltrans District:	7	CA Post Mile:	26.84
County:	Los Angeles	Absolute Post Mile:	50.612
City:	Los Angeles	Length:	0.600
Freeway:	405-S	Latitude:	33.9986
Name:	Braddock	Longitude:	-118.4079

PeMS Website

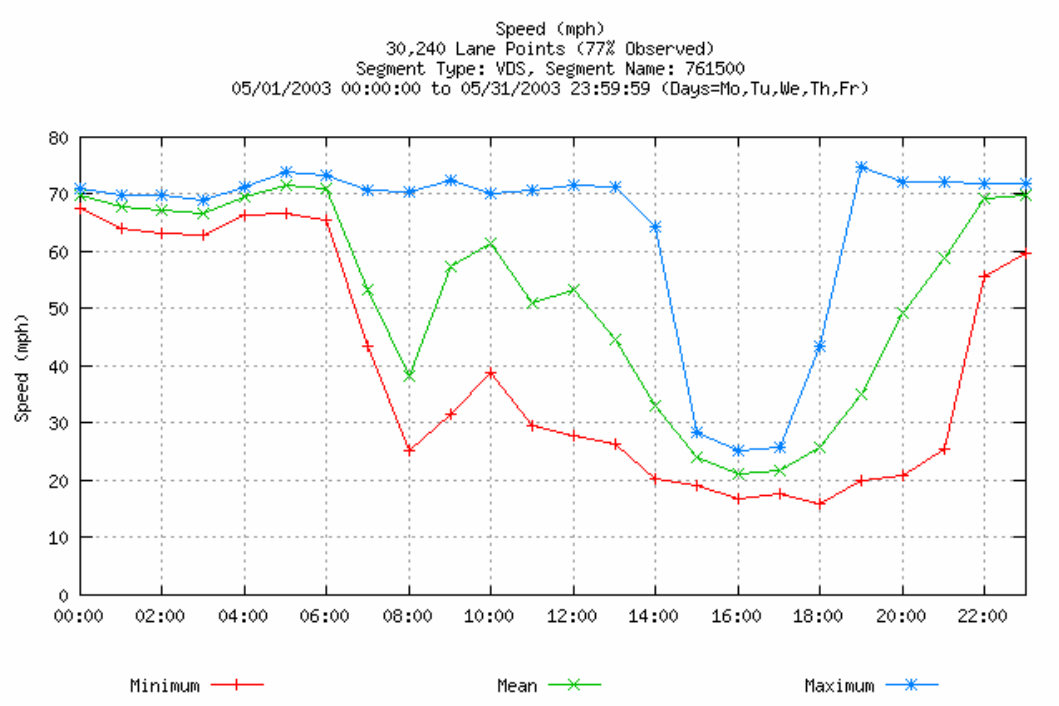


Extracted Dataset (MATLAB)

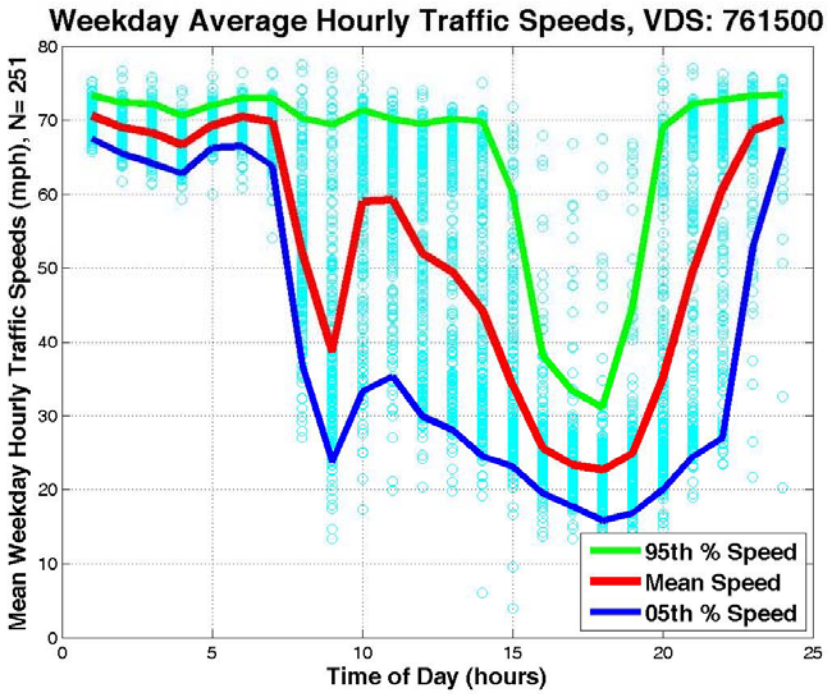


Sample 7 (Speed): Mainline #761500

PeMS Website



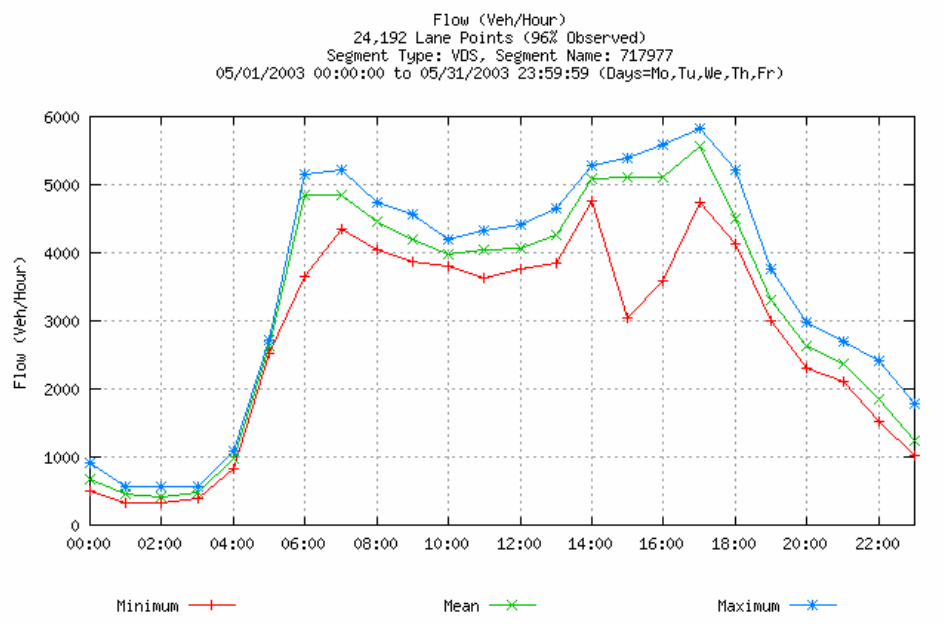
Extracted Dataset (MATLAB)



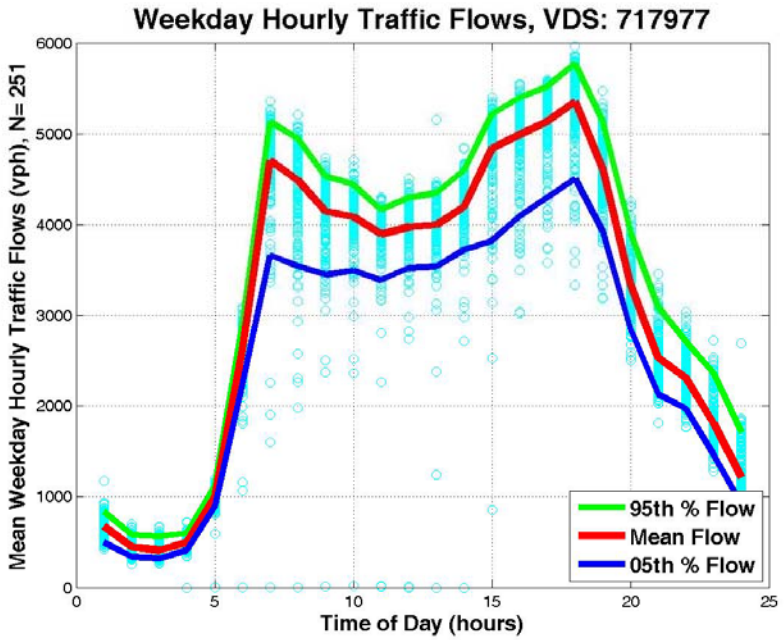
Sample 8 (Flow): Mainline #717977

Caltrans District:	7	CA Post Mile:	R15.9
County:	Los Angeles	Absolute Post Mile:	10.940
City:	Lynwood	Length:	.400
Freeway:	710-N	Latitude:	33.9140
Name:	King 1	Longitude:	-118.1791

PeMS Website

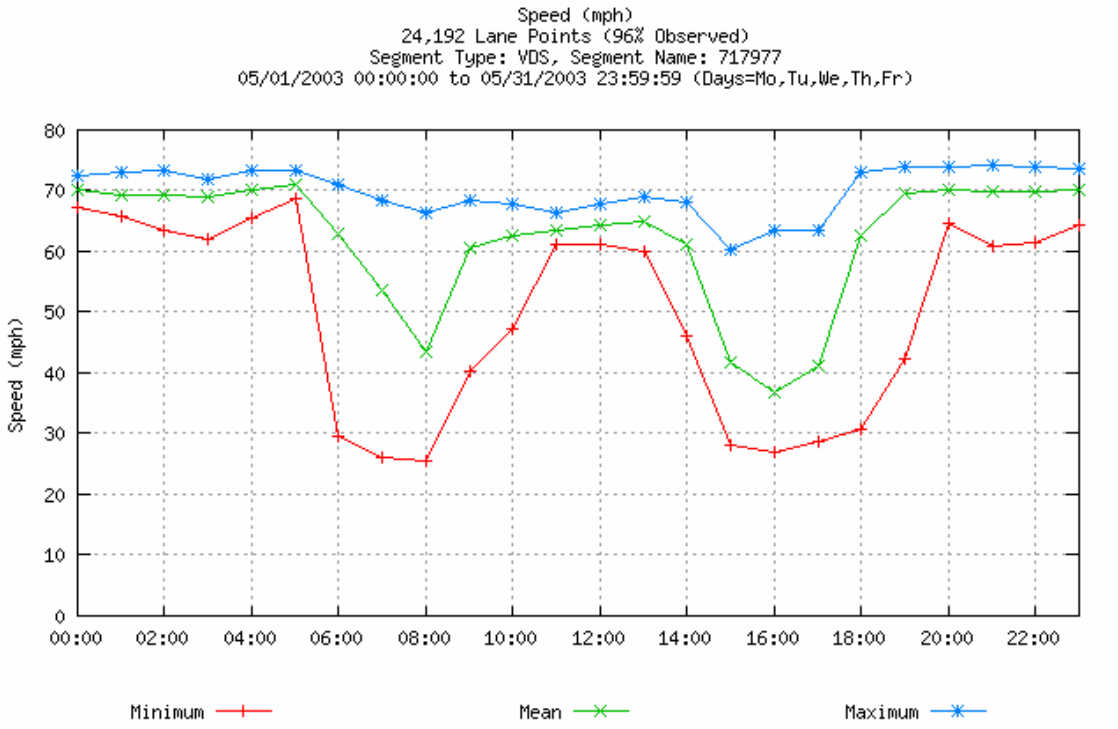


Extracted Dataset (MATLAB)

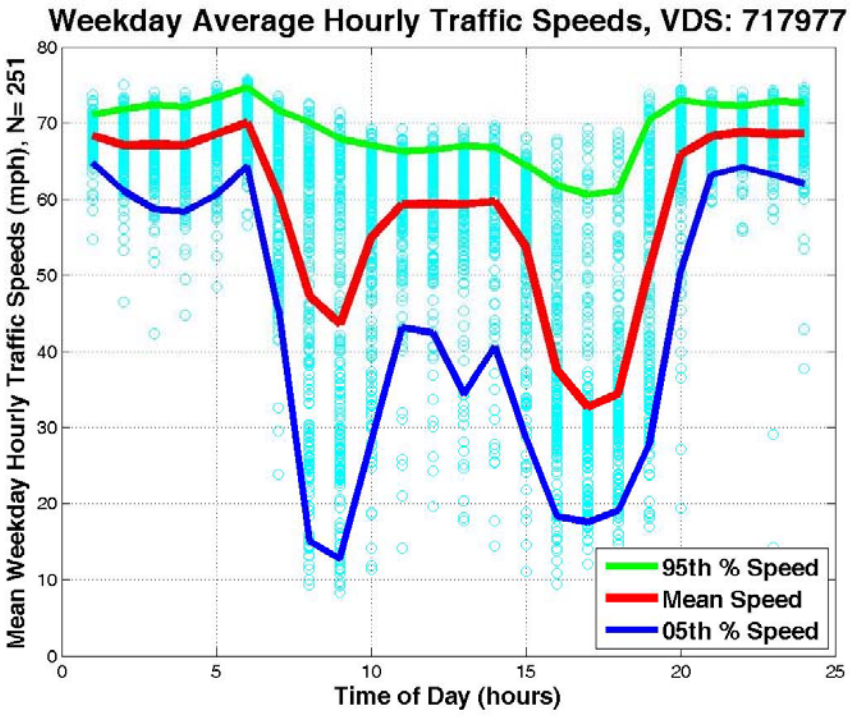


Sample 8 (Speed): Mainline #717977

PeMS Website



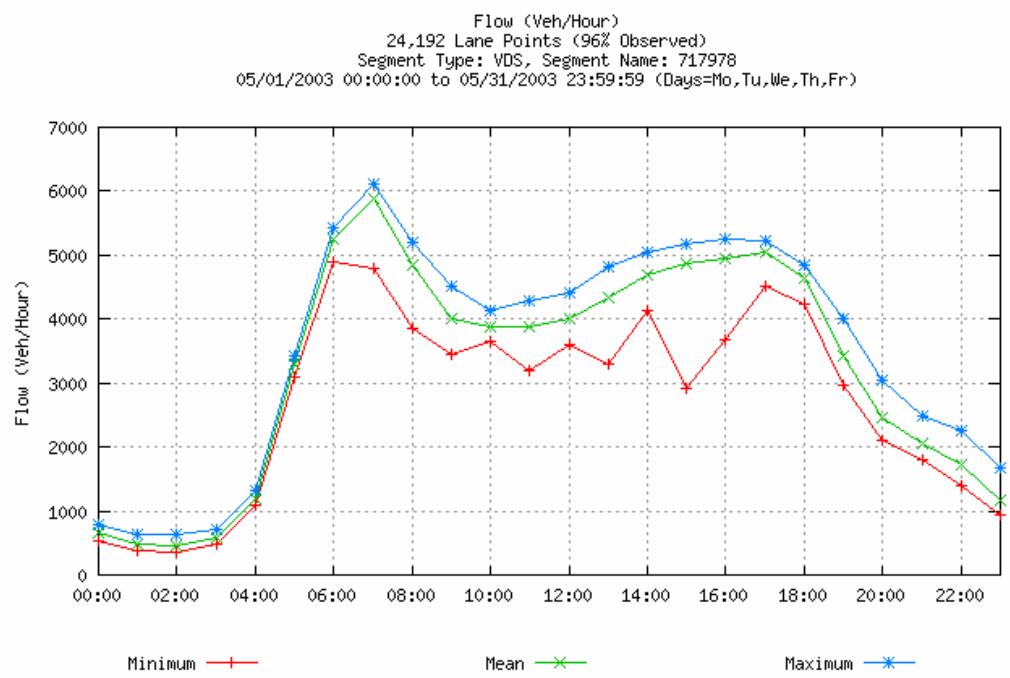
Extracted Dataset (MATLAB)



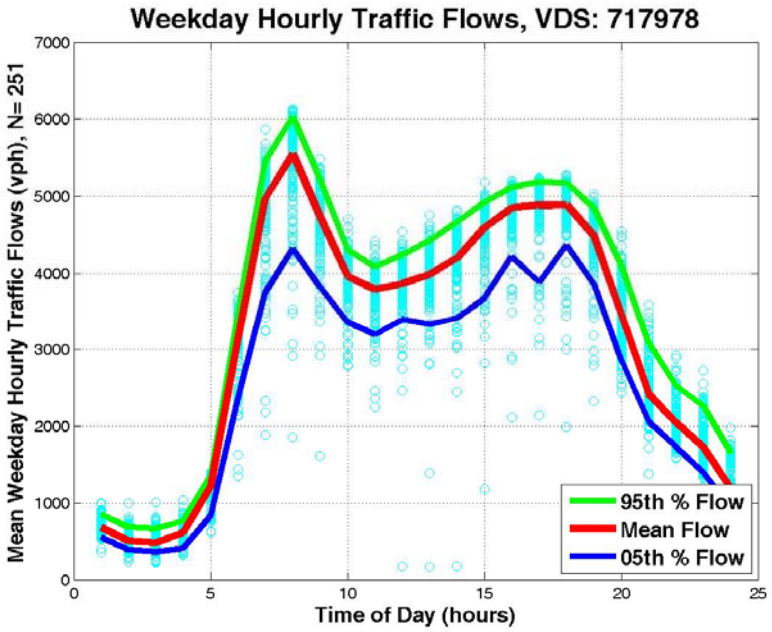
Sample 9 (Flow): Mainline #717978

Caltrans District:	7	CA Post Mile:	R15.9
County:	Los Angeles	Absolute Post Mile:	10.940
City:	Lynwood	Length:	0.793
Freeway:	710-S	Latitude:	33.9140
Name:	King 1	Longitude:	-118.1791

PeMS Website

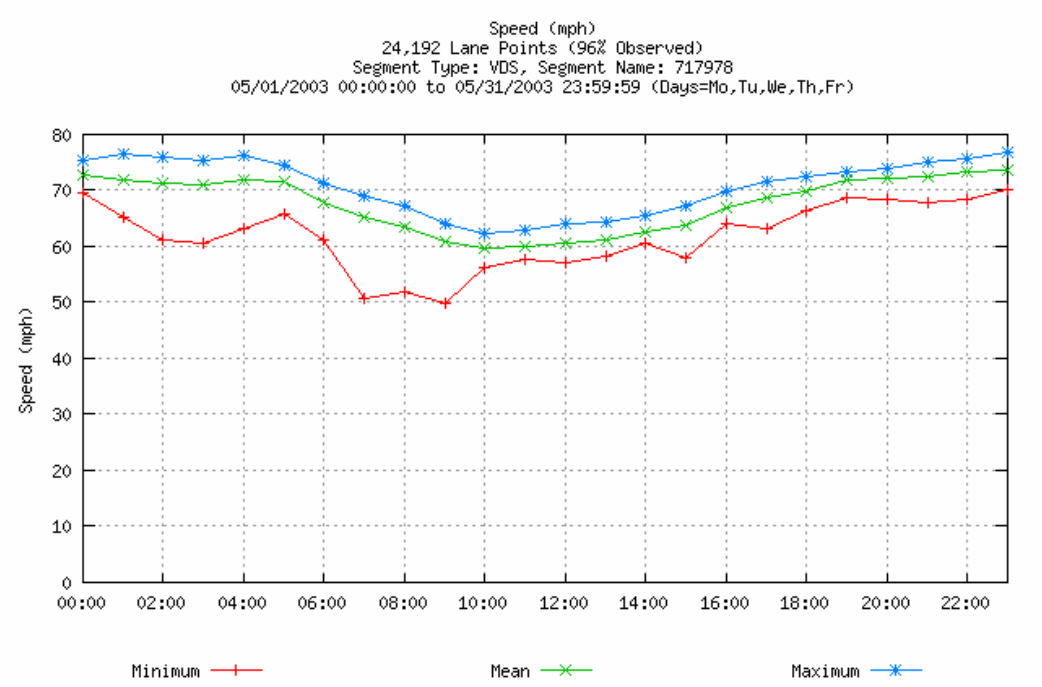


Extracted Dataset (MATLAB)

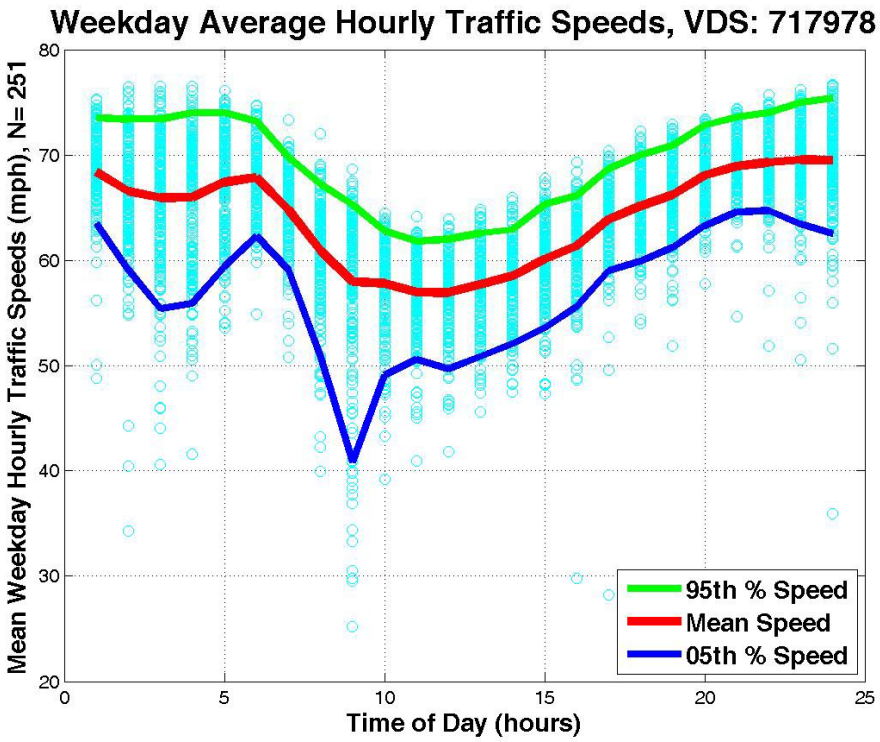


Sample 9 (Speed): Mainline #717978

PeMS Website



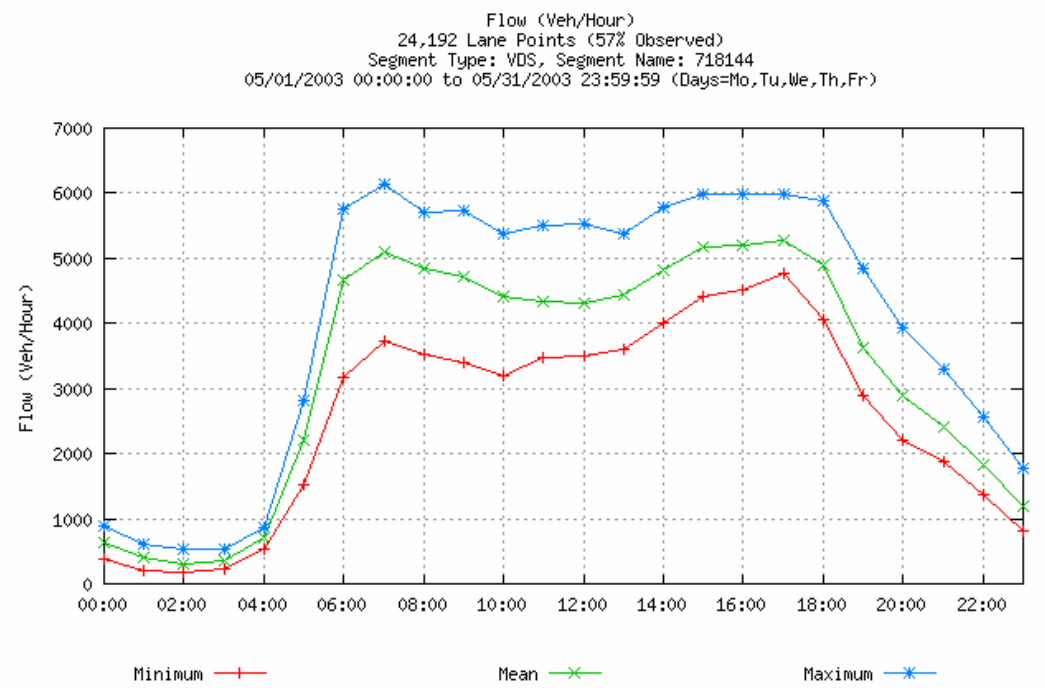
Extracted Dataset (MATLAB)



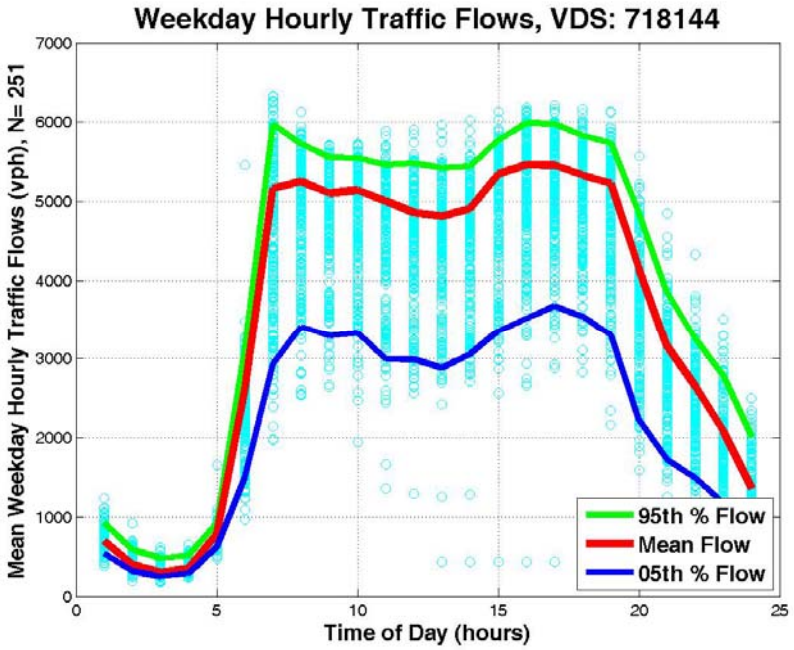
Sample 10 (Flow): Mainline #718144

Caltrans District:	7	CA Post Mile:	25.67
County:	Los Angeles	Absolute Post Mile:	27.003
City:	Los Angeles	Length:	0.935
Freeway:	101-S	Latitude:	34.1696
Name:	Ventura 1	Longitude:	-118.6117

PeMS Website

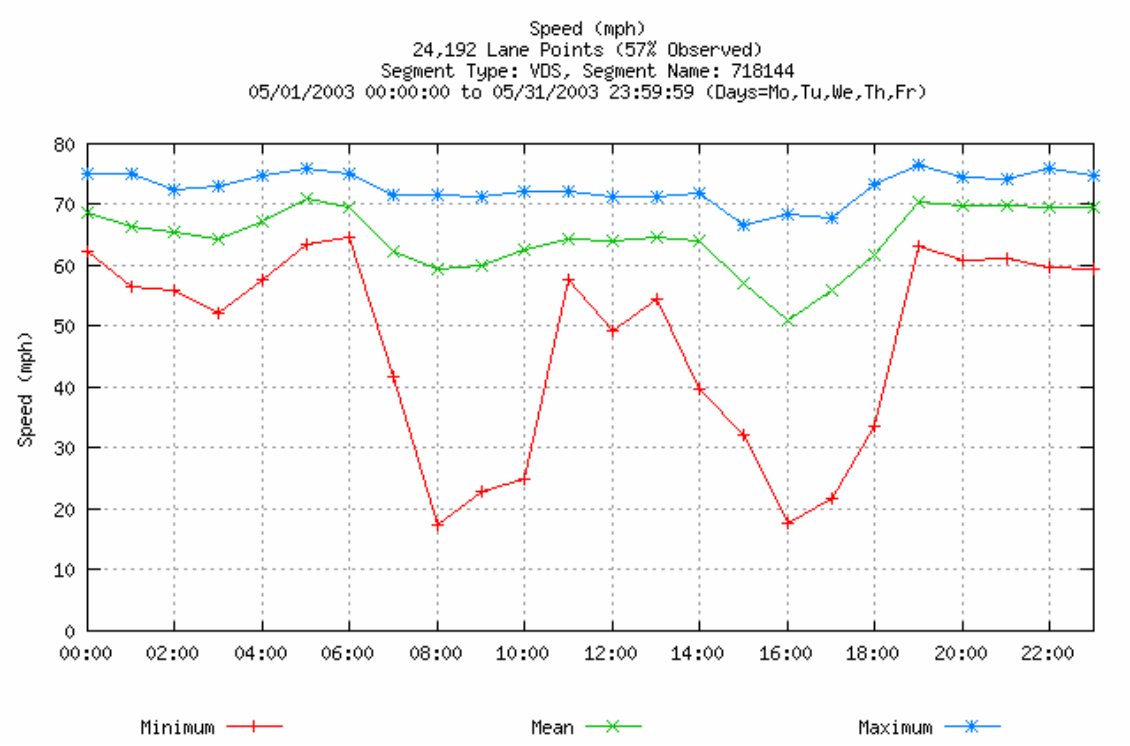


Extracted Dataset (MATLAB)



Sample 10 (Speed): Mainline #718144

PeMS Website



Extracted Dataset (MATLAB)